



# VISDA (Visual Statistical Data Analyzer)

## Cluster Modeling, Visualization, Discovery

Yue Wang, Zuyi Wang, Jianhua Xuan, Yitan Zhu, Robert Clarke

Virginia Polytechnic Institute and State University

The Catholic University of America

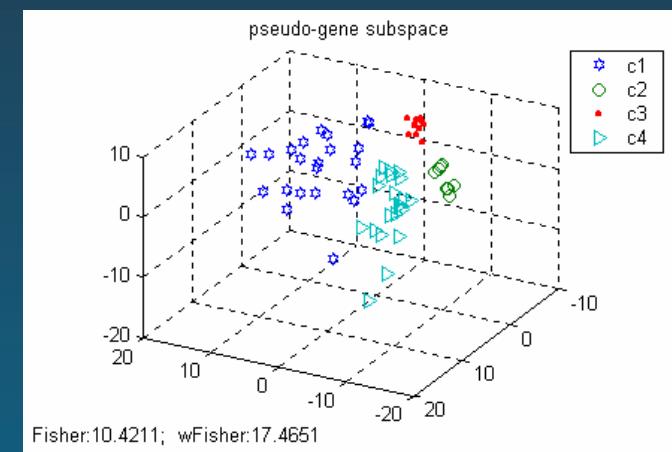
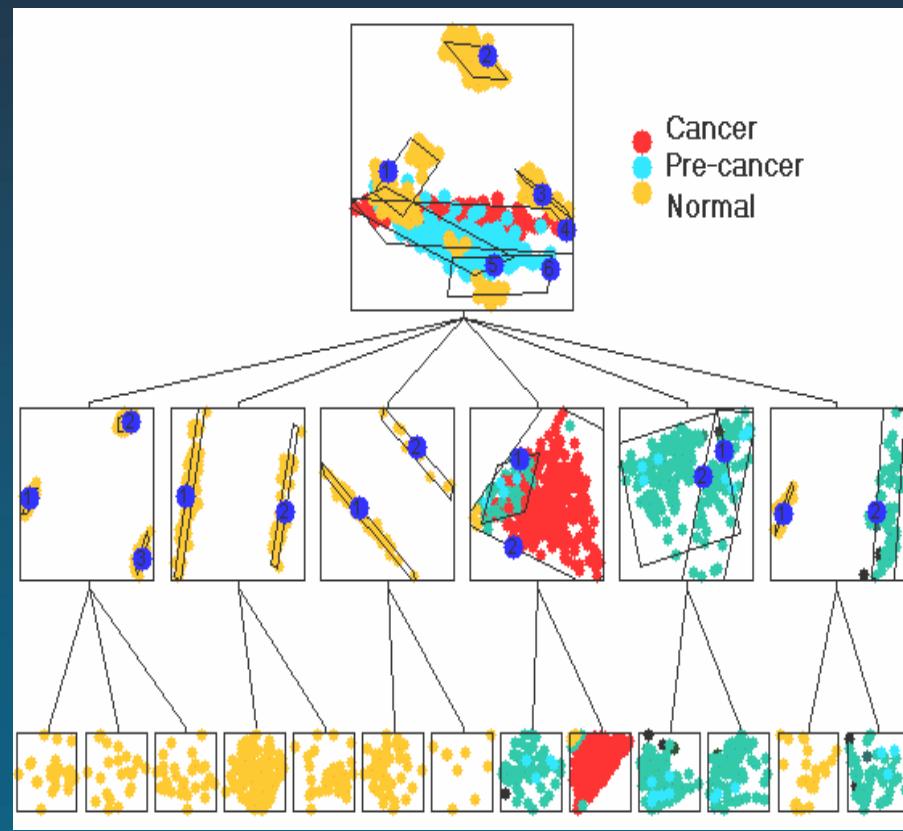
Georgetown University Lombardi Cancer Center



# Motivations

As a step toward understanding multivariate gene expression or other types of biological data sets, **cluster** information reveals insight that may prove useful in knowledge discovery.

Multivariate statistical **visualization** (human-data interaction and exploration) has proven to be a powerful tool for the analysis and interpretation of such complex data sets.





# Challenge & Strategy

The growing volume of biological data sets are often high dimensional, multimodal, and lacking in prior knowledge

Utilize the (nonlinear) power of multivariate higher-order statistics

Incorporate the human gift for pattern recognition



# Principal Methods

Projective discriminative visualization: when there are more than three variables, it stretches the imagination to visualize the data structure

Hierarchical cluster modeling: a single projection of the data onto a visualization space may not be able to capture all of the interesting aspects of the data set.

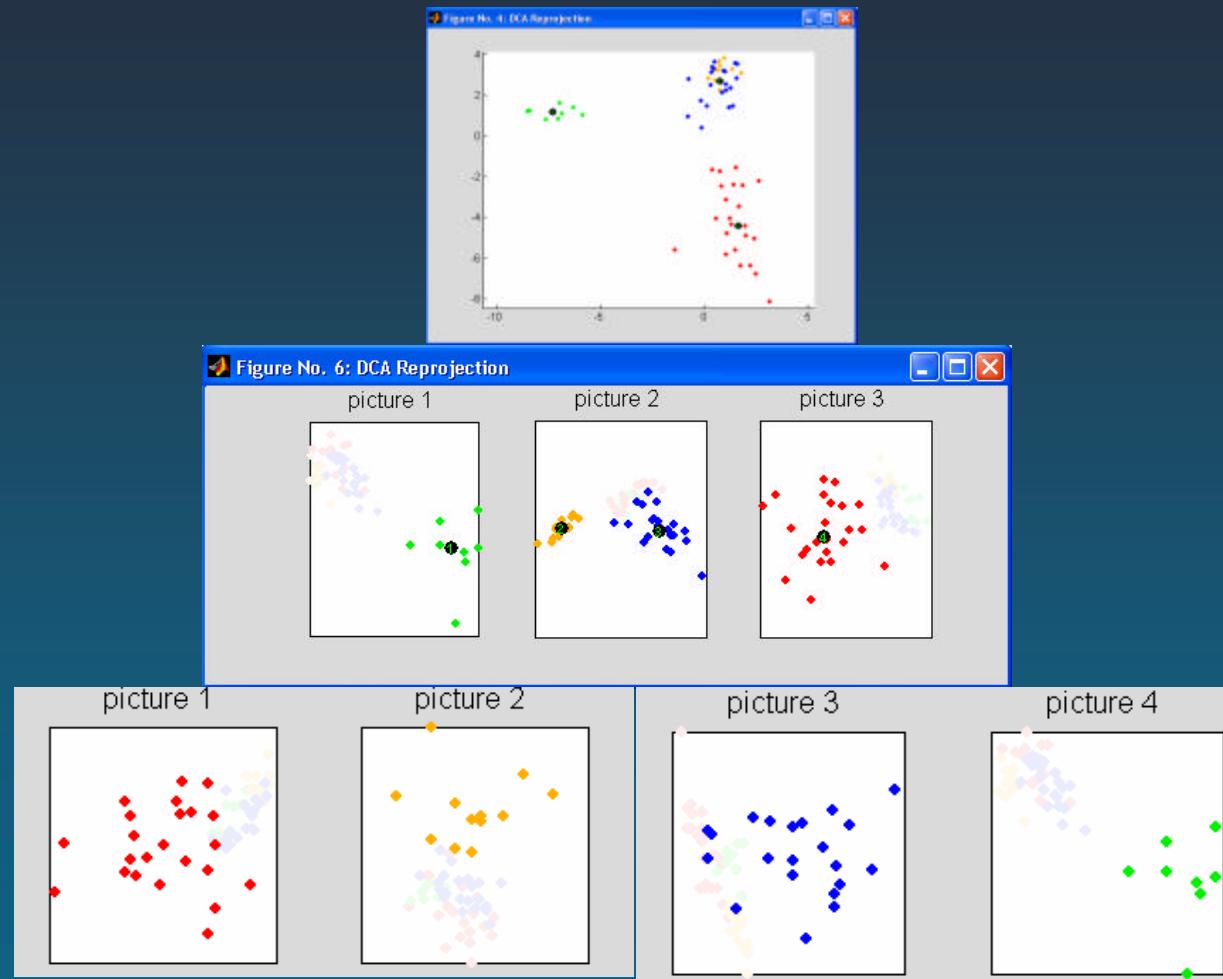


# Integrative Components

Mixture modeling

Soft clustering

Discriminative projection





# Some Existing Approaches

- Data classification (defining clusters):
  - Hierarchical clustering (HC)
  - Self-organizing map (SOM)
- Data projection (visualizing clusters):
  - Principal component analysis (PCA)
  - Multidimensional scaling (MDS)
- Associated major potential **limitations**:
  - Non-statistical, non-robustness, no cluster selection
  - Non-discriminative, single projection, low scalability
  - Disconnection between the two major operations



# Hierarchical Modeling

$$p(\mathbf{t}) = \sum_{k=1}^{K_1} p_k \sum_{j=1}^{K_{2,k}} p_{j|k} g(\mathbf{t} | \boldsymbol{\mu}_{\mathbf{t}(k,j)}, \mathbf{C}_{\mathbf{t}(k,j)})$$

where for each cluster,  $\mathbf{m}$  and  $\mathbf{C}$  are the mean vector and covariance matrix, respectively,  $\pi$  is the relative mass,  $g$  is the gaussian kernel,  $K_1$  is the cluster number identifiable at top level,  $K_{2,k}$  is the cluster number identifiable at second level.

# Projection Portfolio

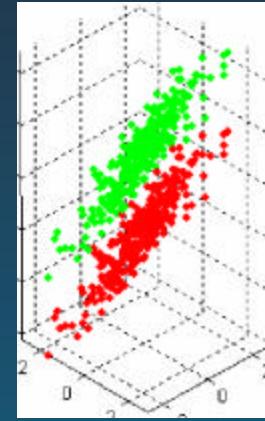
- Discriminatory visualization (separability of patterns):

- Unsupervised mode: projection pursuit method (PCA, PCA-PPM, ICA)

$$nkurt(x_i) = \frac{1}{N} \sum_{j=1}^N \left( \frac{x_{i,j}}{\mathbf{s}_x} \right)^4 - 3$$

- Model-supported mode: HC initiated weighted Fisher criterion (HC-wFC-DCA)

$$J_{wF}(\mathbf{W}) = \sum_{k=1}^{K_0-1} \sum_{l=K+1}^{K_0} \mathbf{p}_k \mathbf{p}_l \mathbf{w}(\Delta_{kl}) trace(\mathbf{W}^T \mathbf{S}_w^{-1} \mathbf{S}_{kl} \mathbf{W})$$





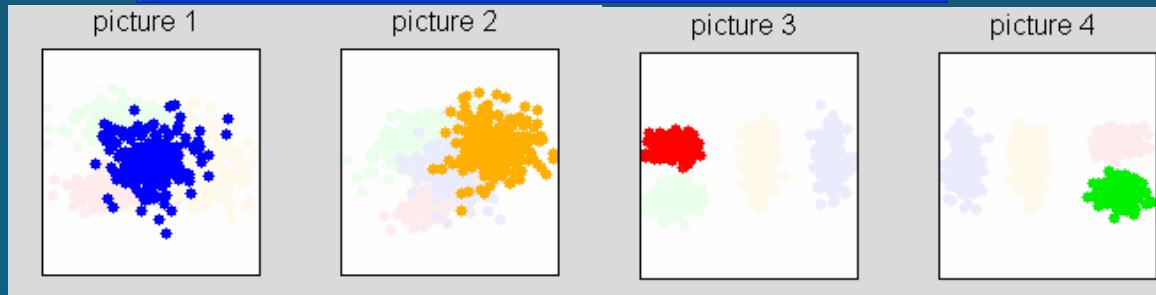
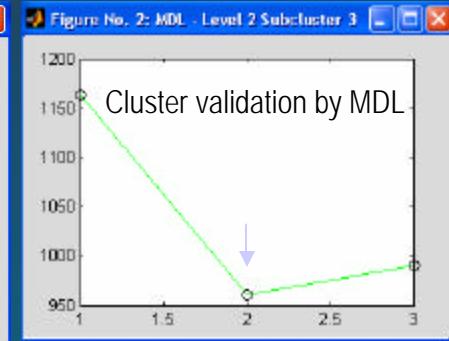
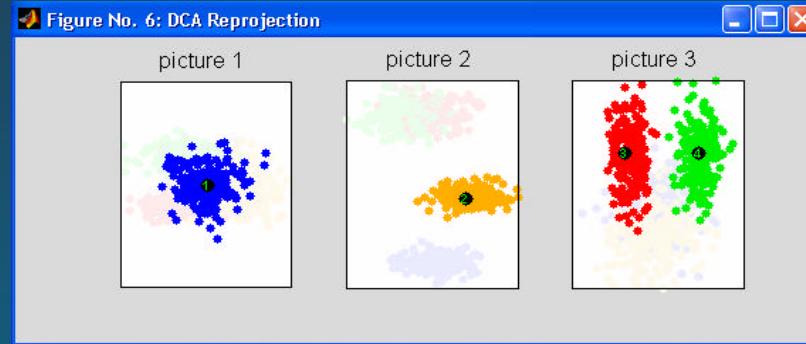
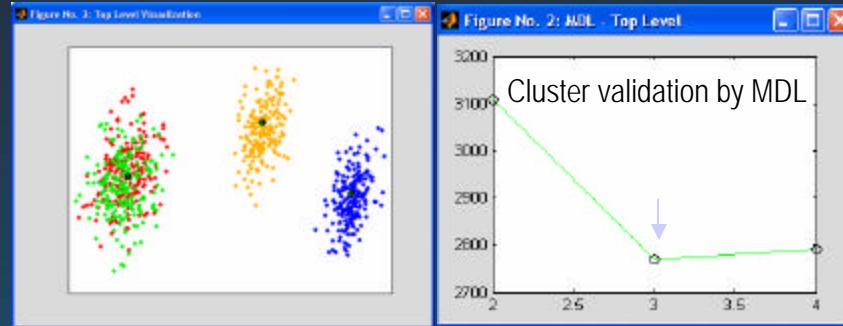
- Bayes soft classification:
  - Hierarchical expectation-maximization (EM) clustering

$$z_{ik} = \frac{\mathbf{p}_k g(\mathbf{t}_i | \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})}{\sum_{m=1}^{K_0} \mathbf{p}_m g(\mathbf{t}_i | \boldsymbol{\mu}_{\mathbf{t}m}, \mathbf{C}_{\mathbf{t}m})}, \quad z_{i(k,j)} = z_{ik} \frac{\mathbf{p}_{j|k} g(\mathbf{t}_i | \boldsymbol{\mu}_{\mathbf{t}(k,j)}, \mathbf{C}_{\mathbf{t}(k,j)})}{\sum_{n=1}^{K_{2,k}} \mathbf{p}_{n|k} g(\mathbf{t}_i | \boldsymbol{\mu}_{\mathbf{t}(k,n)}, \mathbf{C}_{\mathbf{t}(k,n)})}$$

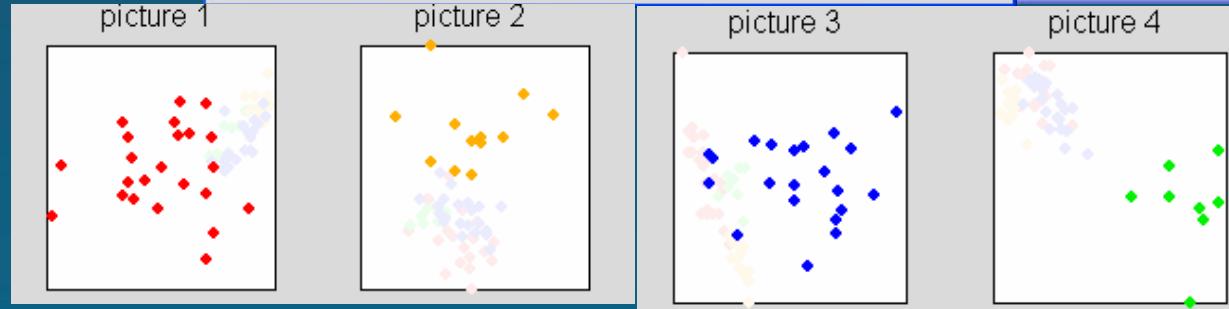
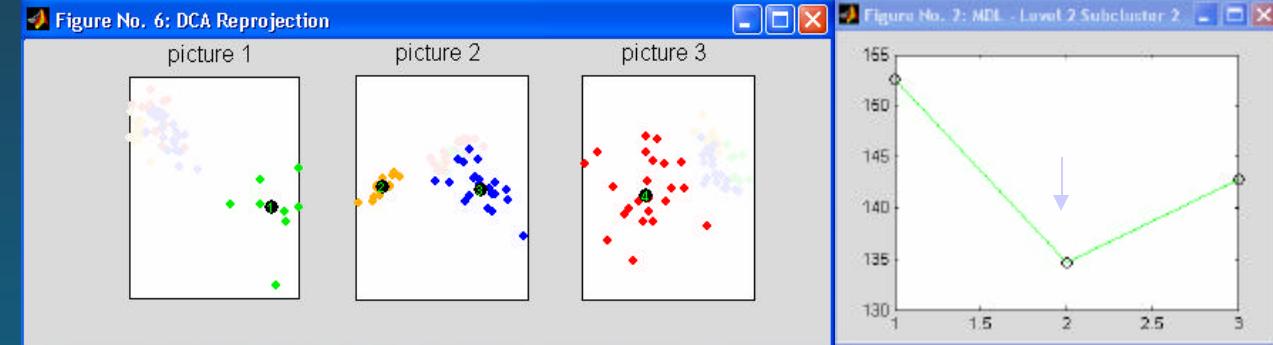
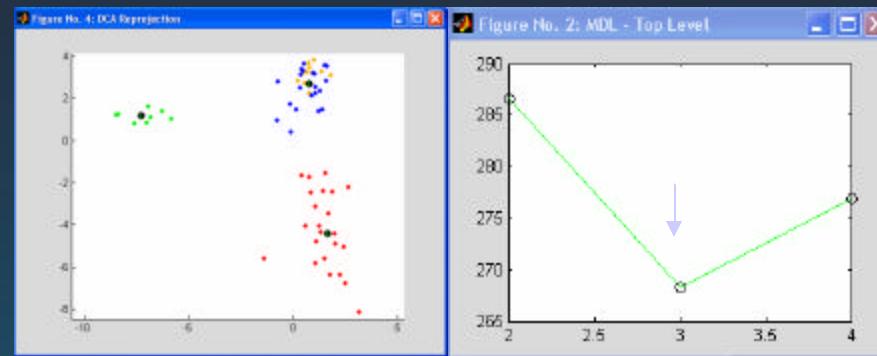
- Model selection by minimum description length (MDL) criterion: minimax entropy principle; bias vs. variance

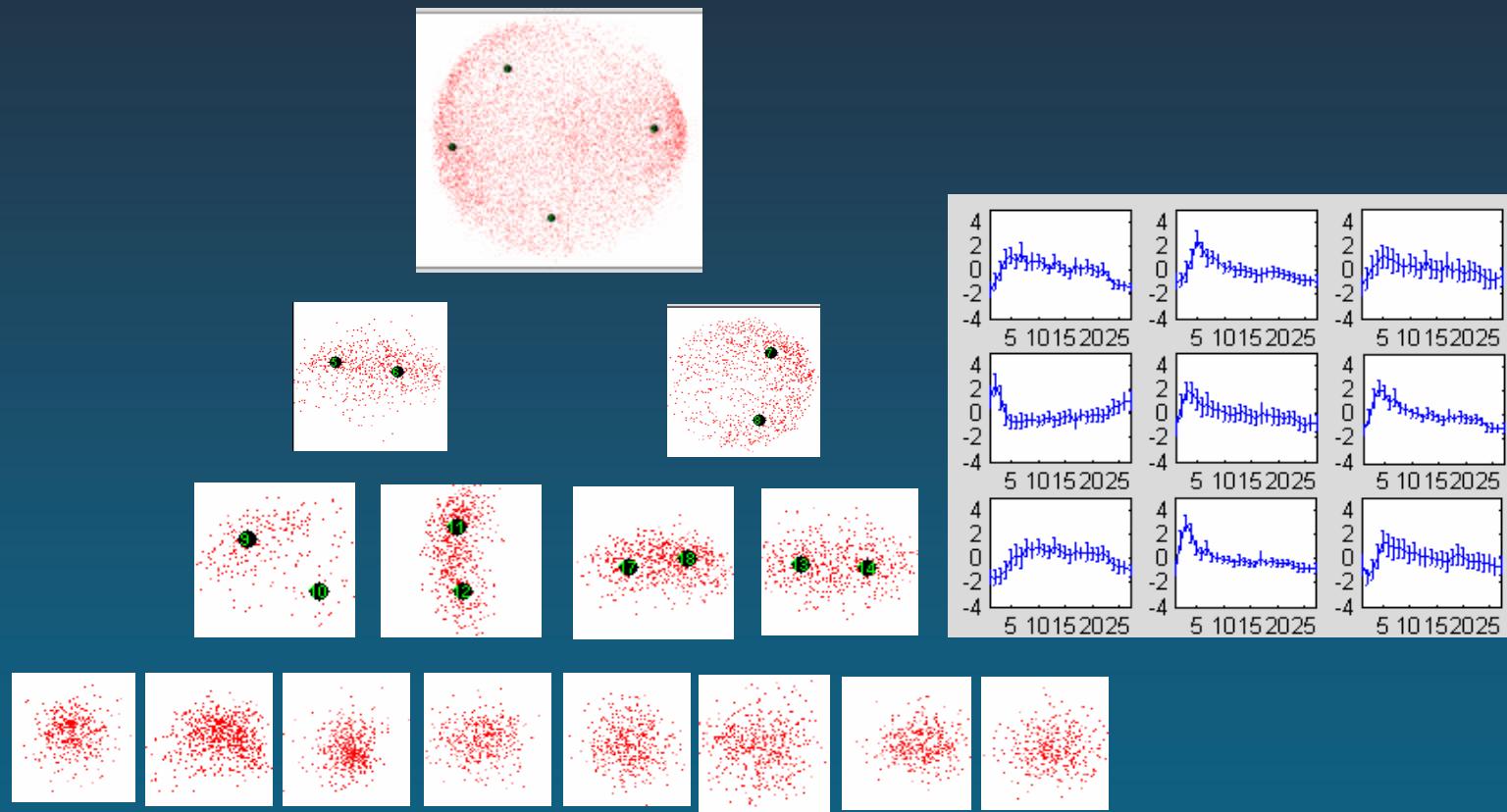
$$MDL(K_0) = - \sum_{i=1}^N \log \left( \sum_{k=1}^{K_0} \mathbf{p}_k g(\mathbf{x}_i | \boldsymbol{\mu}_{\mathbf{x}k}, \mathbf{C}_{\mathbf{x}k}) \right)_{ML} + \frac{6K_0 - 1}{2} \log N$$

# Simulation with Truth



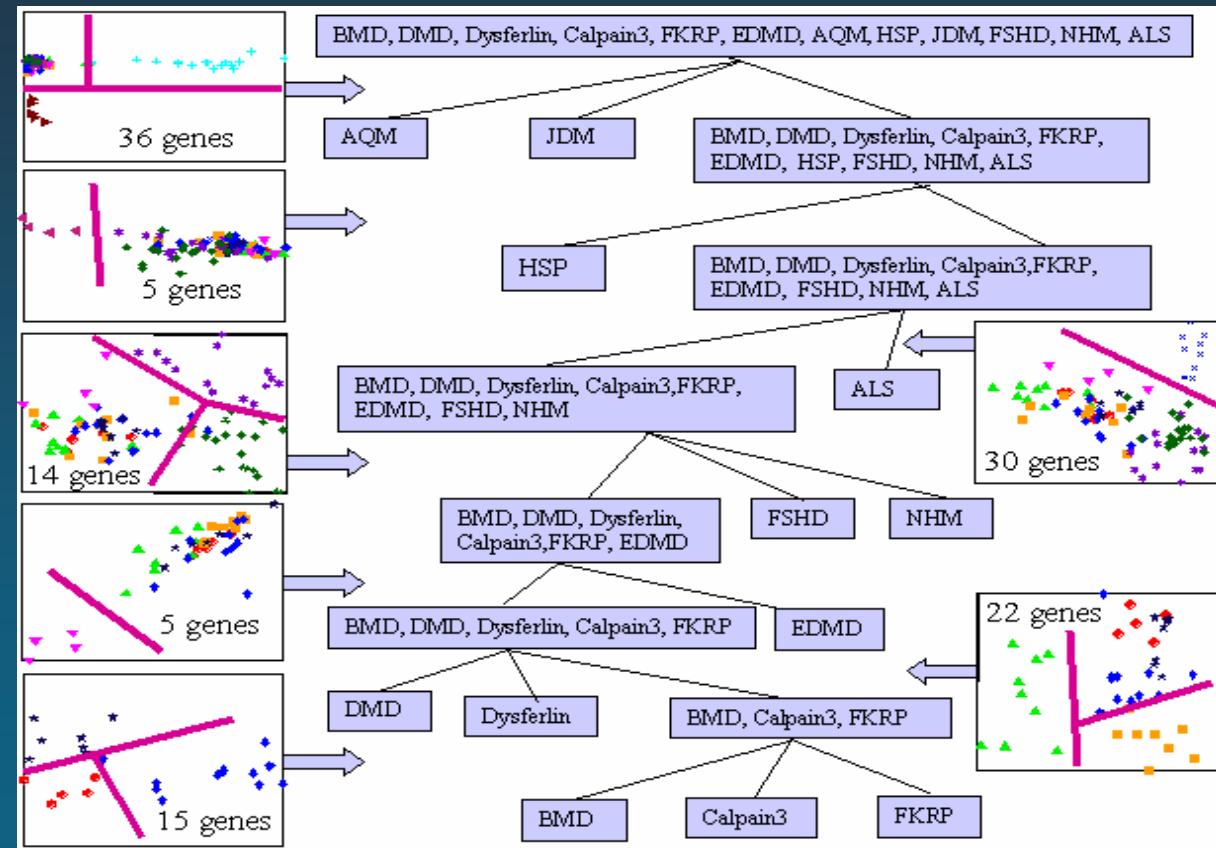
# Real SRBCT with Truth





# Adaptive Subspace Tree

Concurrent feature/gene selection and extraction





# Illustration with ALL/AML

